Not just what, but also when: Guided automatic pronunciation modeling for Broadcast News

Eric Fosler-Lussier (1) and Gethin Williams (2)

(1) International Computer Science Institute
University of California, Berkeley
1947 Center Street, Suite 600
Berkeley, CA 94704

(2) Sheffield University
Department of Computer Science
Sheffield, England

ABSTRACT

In this paper, we describe improvements to the pronunciation model featured in the 1998 SPRACH Broadcast News evaluation system. Various smoothing and pruning techniques and the integration of confidence scores into the pronunciation model training provided a 4% relative improvement over the baseline model. We also report on promising new techniques that did not appear in the evaluation system.

1. INTRODUCTION

A recent surge of efforts in automatic pronunciation modeling within the ASR community has yielded mixed results for large-vocabulary speech recognition systems. Simply adding raw pronunciations from phone recognition to a dictionary can vastly increase decoding time, often with very little benefit. It is therefore important not only to discover *what* alternative pronunciations are possible, but also to introduce extra pronunciations only *when* they are needed. In the work described in this paper we sought to discover ways in which we could improve the baseline lexicon for the SPRACH Broadcast News System [1]. We used the tools of static[1] baseform learning and decision-tree (d-tree) modeling to determine a range of new pronunciation alternatives. Posterior probability-based acoustic confidence measures derived from the system's connectionist acoustic model and new pruning techniques guided our selection of baseforms.

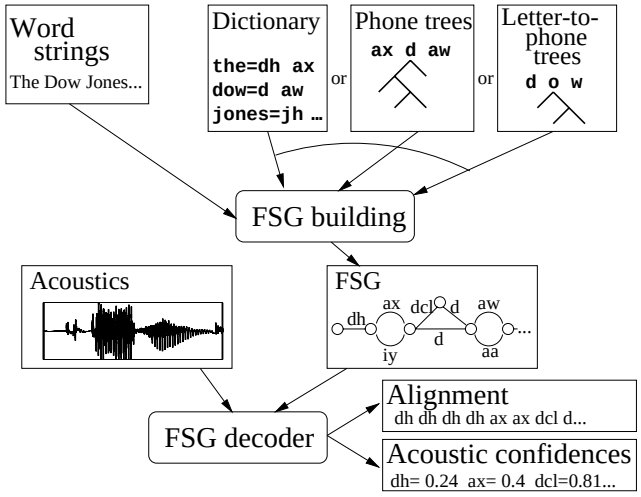

Figure 1: Encapsulation of decoding for different pronunciation models using finite state grammars.

[1] We contrast *static* pronunciation models with those that are sensitive to contextual factors such as speaking rate.

This year's evaluation brought several new challenges in pronunciation modeling. Our primary focus was to apply lessons learned from automatic pronunciation modeling work in the Switchboard domain [7, *inter alia*], increasing the diversity of pronunciations in the dictionary. We were also faced with the issue of novel words— a vocabulary that spans current affairs is crucial for Broadcast News (BN) recognition. For words that did not occur previously in our dictionary, like *Lewinsky*, we needed to generate new baseforms quickly.

Each of these tasks required considerable new machinery in our training system. In order to save effort in implementing these disparate functions, we reorganized our pronunciation software around a finite state grammar (FSG) decoder (Figure 1). The modularization of Viterbi alignments into an FSG compilation stage and a decoding stage allowed for novel compilation techniques without our having to completely rewrite the decoder. Thus, we could easily implement new pronunciation models (such as a letter-to-phone model for new words) as long as the output of the procedure was a valid finite state grammar. The decoder can also be used to rescore n-best lists, facilitating the use of complex pronunciation models.

One of the benefits of using a connectionist acoustic model in conjunction with our FSG decoder is the availability of posterior probability-based acoustic confidence measures at both the phone and the word levels [8]. We found these acoustic confidence measures useful on several occasions as a guide to model building, including selection of pronunciations and checking automatically created models for novel words.

Since the acoustic model was being developed at the same time as the pronunciation models, we had the option of using either an acoustic model that was not as good as the evaluation model or a shifting acoustic model baseline. We opted for the former; as a consequence, many of the word error rate (WER) results reported here are 3-4% higher than those obtained using our final evaluation system. We combined the 1997 ABBOT PLP-based recurrent neural network (RNN) context-independent phone classifier [3] with a 4000 hidden unit multi-layer perceptron (MLP) using modulation-filtered spectrogram features [6]. Both networks were trained only on the 1997 BN training data. For fast turnaround development testing, we used a half-hour subset of the 1997 evaluation set (labeled Hub-4E-97-subset), which also increased error rates by roughly 2% over the full 1997 evaluation set.

2. GENERATION OF NEW PRONUNCIATIONS

The baseline lexicon was derived from the 1996 ABBOT Broadcast News transcription system [2] and contained an average of 1.10 pronunciations per word for a 65K vocabulary. In order to increase the

Lexicon	Decoding Parameters	
	7 Hyp. WER (%)	27 Hyp. WER (%)
Baseline: ABBOT 96	29.9	27.5
Augmented: λ=0.5	28.9	27.1

Table 1: Word error rate on Hub-4E-97-subset for static lexica.

number of pronunciations available to the recognizer, our first task was to align the *canonical* transcription of the acoustic training data against an *alternative* transcription. The canonical transcription was obtained from a forced Viterbi alignment of the reference word sequence to the training data using the baseline lexicon, whereas the alternative transcription was obtained by running the recognizer using only the phone-level constraint of a phone bigram. Each transcription covered the 100-hour 1997 training set.

In the second stage of the pronunciation model learning process, we trained d-trees to predict the phone-recognition realization of a dictionary phone using the alignment between the canonical and alternative transcriptions. D-trees estimated a probability distribution over the realization of dictionary phones using the identity, manner, place, and syllabic position of each phone and its immediate neighbors as contextual features. 90% of the 1997 BN training data was used for training the d-trees, and 10% for tree pruning. The distributions from the d-trees were then compiled into an FSG: for the nth phone in the canonical transcription, the appropriate tree distribution d was found. Between nodes n and $n+1$ in the FSG, an arc was added for every recognition phone in d, appended with the appropriate probability. Phone deletions were accommodated through the insertion of null transitions. Some smoothing was applied to this FSG construction by disallowing any transitions with below threshold probabilities (the threshold was arbitrarily set to 0.1).

Following d-tree training and FSG compilation, in the third stage we created a new static lexicon. The compiled FSG was realigned to the training data to obtain a *smoothed* phone-constraint decoding in the spirit of Riley *et al.* [7] (although they used hand-transcriptions as a starting point, rather than phone recognition). Since the FSG decoder produced both a word and phone alignment, the new alternative transcription was easily converted into a new static lexicon for our full first-pass decoder (NOWAY). However, we found that the resulting lexicon was still too noisy, particularly for infrequently occurring words. We therefore merged the newly obtained pronunciations with those from the baseline ABBOT 96 lexicon, using the following interpolation:

$$P_{\mathcal{L}}(\text{pron}|\text{word}) = \lambda P_{\mathcal{L}_{\text{new}}}(\text{pron}|\text{word}) + (1-\lambda) P_{\mathcal{L}_{\text{ABBOT}}}(\text{pron}|\text{word})$$

The value of the empirically determined smoothing parameter λ did not affect results much within a broad range of values, so we set $\lambda = (1-\lambda) = 0.5$. Since the weighting factor can be interpreted as a measure of trust in the source of a word's baseforms, a possible strategy would be to make λ dependent upon frequency of a word's occurrence in the training data, although we have not tried this.

In a narrow pruning beam width decoding (Table 1:7 hypothesis decoding), the augmented dictionary outperformed the ABBOT 96 dictionary. When a wider beam width was used (max 27 hypotheses) the augmented lexicon still provided a gain, but by a smaller margin.

Lexicon	Pruning Style	% WER	Timing
Baseline	n/a	29.9	1.81× RT
Augmented	no pruning	28.9	6.69× RT
prune low probability prons	$p_{\text{pron}} < 0.1 * p_{\max}$	29.5	2.50× RT
	$p_{\text{pron}} < 1.0 * p_{\max}$	31.4	1.85× RT
Count-based pruning	log count α =1.2	28.8	2.72× RT

Table 2: Word error rate on Hub-4E-97-subset for various pruning methods using narrow (7 Hypothesis) decoding parameters.

3. DICTIONARY PRUNING

The new dictionary described above increased the number of pronunciations per word from 1.10 to 1.67, but decoding time increased almost four fold, which was devastating for our 10× real time system. In order to reduce the number of pronunciations and hence the decoding time, we investigated two dictionary pruning techniques. In our traditional pruning scheme, baseforms were removed from the lexicon if they had a prior probability that was less than some fraction of $p_{\max}$, the prior of most probable baseform for the word. While this significantly reduced decoding time, it also halved the gains from the new dictionary, as shown in Table 2. Reducing the lexicon to a single baseform per word (pruning level 1.0) also significantly hurt performance with no corresponding speedup relative to the baseline.

Since high-frequency words usually have more pronunciation variants in continuous speech, we developed a new pruning technique based on the number of occurrences of the word in the training data. In this second scheme, the maximum number of baseforms per word w_i was determined by

$$\#\,\text{baseforms}(w_i) = \alpha \log_{10} \text{count}(w_i) \ ,$$

where α is a tunable parameter to shift the log scaling. The n most likely baseforms for each word included in the lexicon. As shown in the bottom section of Table 2, this method facilitated lower decoding times (only 1.5 times that taken by the baseline) without any increase in WER relative to the unpruned lexicon. The results in Table 3 show that gains provided by the log-count pruning scheme carry over to the wider beam decoding condition. A lexicon pruned using this second scheme was therefore selected for use in the SPRACH 98 system; we found that the modest improvements from this lexicon were duplicated across test sets (including the full 1997 Hub4 Evaluation) and with different acoustic models.

Following the evaluation, we computed the posterior probability based average acoustic confidence scores for the baseforms in the unpruned lexicon from a forced Viterbi alignment to the 1997 BN training set. Baseforms were reselected using the log-count pruning

Lexicon	WER (%)	Decode time
Baseline	27.5	21.73 x RT
New: no pruning	27.1	72.03 x RT
log count (SPRACH 98)	26.9	33.07 x RT
confidence log count	26.6	30.45 x RT

Table 3: Word error rate on Hub-4E-97-subset for various pruning methods using full (27 Hypothesis) decoding parameters.

scheme according to their confidence-based rankings; this provided a small boost to performance both in terms of decoding time and recognizer accuracy.

4. CONFIDENCE-BASED EVALUATION OF NOVEL WORD PRONUNCIATIONS

As indicated in the introduction, one problem we encountered was determining pronunciations for novel words not already in our dictionary that occurred either in the training transcriptions or in language model training texts. Within our pronunciation software framework, this involved construction FSGs directly from the orthography of the word, using acoustic alignment to determine the best pronunciation.

Building models to predict the pronunciation of a word from its orthography required two steps: (1) we aligned the letters in the dictionary to corresponding phones using a hidden Markov model; and (2) we trained letter-to-sound d-trees to estimate the probability distribution over phones given a central letter and the context of three letters to the left and three letters to the right.

Given a set of letter-to-sound trees, it was then possible to construct a (bushy) pronunciation graph for a novel word, and align this graph to acoustics using the FSG decoder. We view the matching of this graph to the acoustic models as the critical gain of this technique; using a text-to-speech system that was uninfluenced by our acoustic models would likely produce pronunciations with different properties than those in our baseline dictionary.

The FSG alignment could only be performed on words for which we had sample acoustics. Therefore, we recorded subjects reading aloud from word lists presented by the computer for several thousand novel words present only in the language-model training texts. The Viterbi alignment of the graph to the acoustics provided both a putative baseform and also an acoustic confidence score. Using this procedure, pronunciations for 7,000 novel words were incorporated into the 1998 SPRACH system. While the procedure was far from perfect, spot checks of the high-confidence novel baseforms showed them to be more reliable than the low-confidence ones. We therefore focused hand correction efforts on lower confidence pronunciations.

5. MULTI-WORD AND DYNAMIC DICTIONARIES

Since the pronunciation of a word is dependent upon contextual factors such as the words that follow and precede it, word predictability and speaking rate, we also investigated ways to add more contextual influence into the pronunciation model.

5.1. Multi-word pronunciations

Our initial attempt at incorporating context was the creation of multi-word baseforms. We elected to create baseforms for the approximately 4,000 word-pairs that occurred sufficiently frequently in the training data to facilitate reliable baseform learning (*i.e.*, those pairs with 20 or more examples). Of these 4k pairs, 500 were selected for inclusion in the lexicon and as single items in the n-gram language model. Three different ranking schemes were investigated:

MW$_{conf}$ Word-pairs were ranked according to their average inverse posterior confidence in a forced Viterbi alignment of the training data.

Lexicon	WER (%) with ABBOT 96	WER (%) with SPRACH 98
Baseline: Dictionary alone	31.9	30.5
+ MW$_{conf}$	32.0	-
+ MW$_{mi}$	32.0	30.5
+ MW$_{mi+freq}$	31.3	30.5

Table 4: Word error rate on Hub-4E-97-subset for multi-word lexica.

MW$_{mi}$ Word-pairs were ranked according to the mutual information between the frequency distributions of the set of observed pronunciations (from the smoothed phone recognition) for the two words (*c.f.* [4]).

MW$_{mi+freq}$ Because MW$_{mi}$ was found to rank some relatively infrequently occurring word-pairs highly, a third scheme was devised that ranked pairs according to both their mutual information and also the frequency of occurrence.

Smaller language models with and without multi-words were built for quick testing purposes, resulting in an increase in baseline error rate. The results from the multi-word experiments (table 4) were inconclusive. When augmenting the baseline ABBOT 96 dictionary, multi-words chosen using the MW$_{mi+freq}$ scheme provided a small improvement. This gain vanished, however, when the same multi-words were incorporated into the SPRACH 98 dictionary.

5.2. Word and syllable-based decision trees

As an enhancement of multi-word pronunciations, we developed d-trees that predicted the pronunciation of words based on the identities of surrounding words. This can be considered an extension of the above multi-word experiment, since the d-tree building techniques used mutual information as the criterion for determining branching splits. An added advantage of d-tree modeling is that other features besides word identity can be used as d-tree features, such as speaking rate and trigram probability, that correlate well with pronunciation changes [5].

We built models for the 550 most frequent words using surrounding word identities, and the identities, manner, place, and syllabic position of neighboring phones as features in the d-tree. We also included information about word length, several estimates of speaking rate, and the trigram probability of the word. Slightly less than half of the trees in each case used a distribution other than the prior (*i.e.*, were grown to more than one leaf).

In building the word trees, we found linguistically plausible pronunciation changes. For example, in the tree for *president* (shown in figure 2), when the following word was *Clinton*, *Clinton's*, or *Boris*, the final /t/ closure was very likely to be deleted. In addition, the velarization of /n/ to [ng] was possible, a likely consequence of the following /k/ in *Clinton('s)*. It is important to note that the velarization requires the deletion of /t/ to be possible; it is easier to learn these co-occurrences when units larger than individual phones are modeled.

In order to increase coverage, we also trained roughly 800 d-trees based on syllable distributions. Each word was given a single canonical syllable transcription, so that words with similar syllabic-internal pronunciation alternations in the ABBOT 96 dictio-

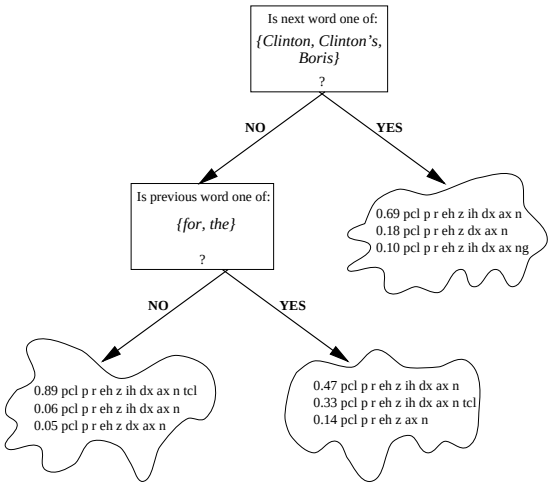

Figure 2: Decision tree for the word "president."

nary shared the same syllable model. In addition to the features found in the word trees, we informed the the syllable trees about the lexical stress of the syllable, position within the word, and the word's identity.

Since the pronunciation scoring required knowledge of the following word in the hypothesis, we were not able to implement these models in our first-pass (NOWAY) decoder. Therefore, we used the dynamic pronunciation model with two decoding strategies: (1) we rescored n-best lists (with $n = 100$) constructed by the NOWAY decoder using our best static models; and (2) we implemented a lattice decoder that re-evaluated word probabilities in the context of a hypothesis. Table 5 summarizes our preliminary results.

In order to test the influence of the decoding process on the results, we recomputed the baseline with the n-best decoder and the lattice decoder using the SPRACH 98 static dictionary. The results in both cases were similar to those of the first-pass decoding (26.9%). The dynamic trees gave us a small (non-significant) increase in accuracy over our improved static lexicon, with syllable trees performing the best. The difference between lattice decoding and n-best rescoring seems to be minimal in this test. We intend to study further the features and models that were most effective in this framework, and the conditions under which they were effective. For example, the 0.4% difference between n-best decoding with the SPRACH 98 dictionary and the syllable trees was accounted for almost completely by a 1.4% improvement in WER in the spontaneous broadcast speech focus condition.

6. CONCLUSIONS

Appropriate smoothing and pruning methods play an important part in building dictionaries for large vocabulary recognition. Particularly when building our $10\times$ real-time system, we found that it is not enough to determine *what* new pronunciations we can install into a new dictionary. One must also consider *when* these pronunciations should be used, either in terms of lexical pruning or determining which pronunciations are appropriate within context.

Lexicon	100-best rescoring	lattice rescoring
Baseline: SPRACH 98	26.7%	27.0%
Word trees	26.5%	26.6%
Syllable trees	26.3%	26.4%

Table 5: Hub4E-97-subset WER for dynamic tree models.

Using decision-tree smoothing of phone recognition to determine what new pronunciations were viable, and a new logarithmic pruning method to decide when to employ these new models, we were able to improve recognition on Broadcast News by about 1% absolute. Confidence measures played a part in identifying which pronunciations matched the recognizer acoustic model, guiding model selection and verification of baseforms for novel words. Finally, contextual methods of determining pronunciations yielded a small improvement in our initial experiments; we feel that more study is needed in this promising area.

7. ACKNOWLEDGMENTS

We are indebted to Gary Cook, Dan Ellis, Adam Janin, and the rest of the SPRACH team for providing acoustic models and the recognition framework for this study. This work was supported by the European Community basic research grant SPRACH, NSF SGER grant IRI-9713346, NSF grant IRI-9712579, and an EPSRC studentship.

References

1. Gary Cook, James Christie, Dan Ellis, Eric Fosler-Lussier, Yoshi Gotoh, Brian Kingsbury, Nelson Morgan, Steve Renals, Tony Robinson, and Gethin Williams. The SPRACH system for the transcription of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, February 1999.
2. G.D. Cook, D.J. Kershaw, J.D.C. Christie, and A.J. Robinson. Transcription of broadcast television and radio news: The 1996 ABBOT system. In *DARPA Speech Recognition Workshop*, Chanitlly, Virginia, February 1997.
3. G.D. Cook and A.J. Robinson. The 1997 ABBOT system for the transcription of broadcast news. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, February 1998.
4. M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Eurospeech-97*, 1997.
5. E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 35–40, Kerkrade, Netherlands, April 1998.
6. Brian E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*. PhD thesis, University of California, Berkeley, California, 1998.
7. M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Kerkrade, Netherlands, April 1998.
8. Gethin Williams. A study of the use and evaluation of confidence measures in automatic speech recognition. Technical Report CS-98-02, Department of Computer Science, University of Sheffield, 1998.